Ordinary Least Squares Regression

Bindeshwar Singh Kushwaha

PostNetwork Academy

Dataset of a Company

Dataset Table:

X (Budget)	Y (Sales)
1	2
2	2.8
3	3.6
4	4.5
5	5.1

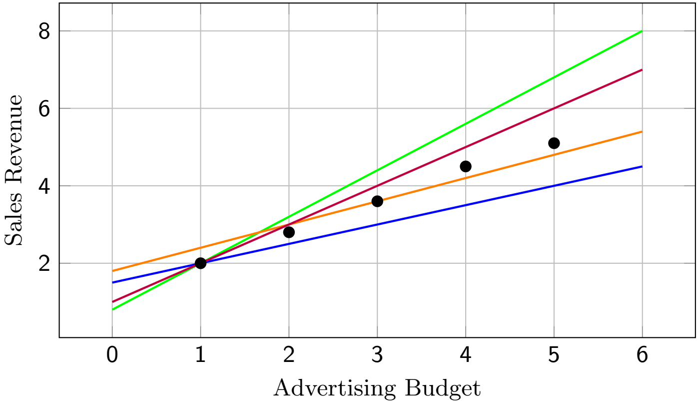
Table: Company's Advertising Budget vs. Sales Data

Description:

- The dataset represents the relationship between advertising budget (X) and sales revenue (Y).
- The company wants to analyze how the budget affects sales using regression analysis.
- The goal is to fit a regression model to predict sales based on the budget.

Multiple Regression Lines Visualization

Visualization of Multiple Lines:



Equations of Different Regression Lines:

- Blue Line: $Y = 0.5X + 1.5$
- Green Line: $Y = 1.2X + 0.8$
- Orange Line: $Y = 0.6X + 1.8$
- Purple Line: $Y = 1.0X + 1.0$

Multiple Regression Lines Visualization

Key Observations:

- There are multiple possible regression lines that can be fitted to the dataset.
- Each line represents a different possible model for predicting sales from the budget.
- We do not know which line is the best fit without an objective criterion like the least squares method.

Ordinary Least Squares (OLS) Regression

Definition: OLS is a statistical method used to estimate the relationship between independent and dependent variables by minimizing the sum of squared residuals.
Objective: Find the best-fit line:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- Y = Dependent variable (response)
- X = Independent variable (predictor)
- β_0 = Intercept (value of Y when $X = 0$)
- β_1 = Slope (change in Y per unit change in X)
- ϵ = Error term (residuals)

Minimizing the Error: The goal is to minimize the sum of squared errors (SSE):

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i$ is the predicted value.
Solution: The OLS estimates for β_1 and β_0 are:

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}, \quad \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Key Idea: The best-fit line minimizes the vertical distance between observed values

Step-by-Step Derivation of β_1 and β_0

Step 1: Define the Sum of Squared Errors (SSE)

$$\mathrm{SSE} = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Step 2: Expand the squared term

$$\mathrm{SSE} = \sum_{i=1}^{n} \left(Y_i^2 - 2Y_i(\beta_0 + \beta_1 X_i) + (\beta_0 + \beta_1 X_i)^2 \right)$$

Step 3: Differentiate SSE with respect to β_0 and β_1 and set to zero

$$\frac{\partial \mathrm{SSE}}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial \mathrm{SSE}}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Step 4: Solve for β_0

$$\beta_0 = \frac{\sum Y_i}{n} - \beta_1 \frac{\sum X_i}{n} = \bar{Y} - \beta_1 \bar{X}$$

Step 5: Solve for β_1

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Data Table for OLS Calculation

X (Budget)	Y (Sales)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
1	2	-2	-1.42	2.84	4.00
2	2.8	-1	-0.62	0.62	1.00
3	3.6	0	0.18	0.00	0.00
4	4.5	1	1.08	1.08	1.00
5	5.1	2	1.68	3.36	4.00
Sum	**18**	0	0	**7.90**	**10.00**

Table: Values required for OLS calculations

Calculation of β_1 and β_0

Step 1: Compute β_1

$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\beta_1 = \frac{7.90}{10.00} = 0.79$$

Step 2: Compute β_0

$$\beta_0 = \bar{Y} - \beta_1\bar{X}$$

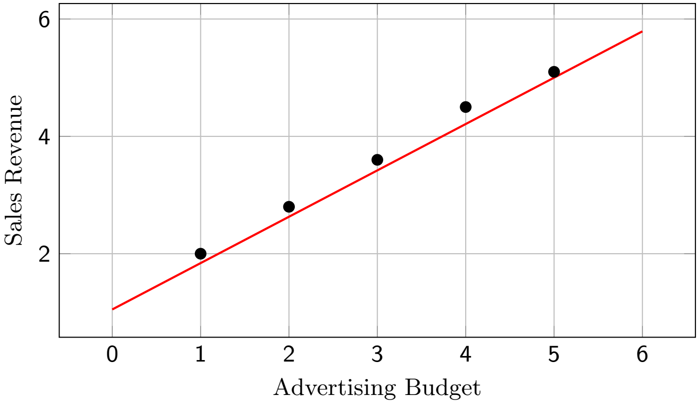
$$\bar{X} = \frac{1+2+3+4+5}{5} = 3, \quad \bar{Y} = \frac{2+2.8+3.6+4.5+5.1}{5} = 3.42$$

$$\beta_0 = 3.42 - (0.79 \times 3) = 3.42 - 2.37 = 1.05$$

Final Regression Equation:

$$Y = 1.05 + 0.79X$$

Best Fit Line Visualization



- The red line represents the best-fit regression line using OLS.
- It is represented by the equation $Y = 0.79X + 1.05$.
- This line minimizes the sum of squared residuals, making it the optimal solution.

Prediction on Unseen Data using $Y = 0.79X + 1.05$

Step 1: Given the Regression Equation

$$Y = 0.79X + 1.05$$

Step 2: Choose an Unseen Value of X
Suppose we want to predict sales (Y) for an advertising budget of $X = 6$.
Step 3: Substitute $X = 6$ into the Equation

$$Y = 0.79(6) + 1.05$$

Step 4: Compute the Predicted Value

$$Y = 4.74 + 1.05 = 5.79$$

Step 5: Interpretation

- If the company spends $X = 6$ on advertising, it is expected to generate $Y = 5.79$ in sales.
- This prediction is based on past trends observed in the dataset.
- The closer the test data is to the training data, the more reliable the prediction.

Reach PostNetwork Academy

Website

www.postnetwork.co

YouTube Channel

www.youtube.com/@postnetworkacademy

Facebook Page

www.facebook.com/postnetworkacademy

LinkedIn Page

www.linkedin.com/company/postnetworkacademy

Thank You!